In: P. Hammer, M. Alirezaie, C. Strannegård (eds.), Proc. Artificial General Intelligence 2023, 286-295

Explicit Goal-Driven Autonomous Self-Explanation Generation

Kristinn R. Thórisson,[1,2] Hjörleifur Rörbeck[1,2]
Jeff Thompson[2] & Hugo Latapie[3]

[1] Reykjavik University, Center for Analysis & Design of Intelligent Agents, Menntavegur 1, Reykjavík, Iceland thorisson@ru.is, hjorleifurh@gmail.com
[2] Icelandic Institute for Intelligent Machines, Reykjavík, Iceland jeff@iiim.is
[3] Cisco Systems, Emerging Technologies & Incubation, San Jose, California, USA hlatapie@cisco.com

Abstract. Explanation can form the basis, in any lawfully behaving environment, of plans, summaries, justifications, analysis and predictions, and serve as a method for probing their validity. For systems with general intelligence, an equally important reason to generate explanations is for directing cumulative knowledge acquisition: Lest they be born knowing everything, a general machine intelligence must be able to handle novelty. This can only be accomplished through a systematic logical analysis of how, in the face of novelty, effective control is achieved and maintained—in other words, through the systematic *explanation of experience*. Explanation generation is thus a requirement for more powerful AI systems, not only for their owners (to verify proper knowledge and operation) but for the AI itself—to leverage its existing knowledge when learning something new. In either case, assigning the automatic generation of explanation to the system itself seems sensible, and quite possibly unavoidable. In this paper we argue that the quality of an agent's explanation generation mechanism is based on how well it fulfils three goals – or purposes – of explanation production: Uncovering unknown or hidden patterns, highlighting or identifying relevant causal chains, and identifying incorrect background assumptions. We present the arguments behind this conclusion and briefly describe an implemented self-explaining system, AERA (Autocatlytic Endogenous Reflective Architecture), capable of *goal-directed self-explanation:* Autonomously explaining its own behavior as well as its acquired knowledge of tasks and environment.

Keywords: Artificial Intelligence · Explanation Generation · Autonomy · General Machine Intelligence · Causal Reasoning · Self-Explanation

1 Introduction

Explainability is an important feature of any artificial intelligence (AI) systems, for numerous reasons. Explanations can form the basis of valid plans, summaries, justifications, predictions, etc. and serve as a method for probing their validity, cost, and potential dangers—which, in fact, is the role of explanations in general

in society. The more complex an AI system is, the more important it is that its operation be transparent and understandable not only by its owners, but also by the system itself. Being explainable implies support for direct inquiries for why a system did what it did, what it plans to do, and why it chose some action over another. The level of transparency offered this way will impact a system's trustworthiness. For any general machine intelligence, trustworthiness is a necessity, since such systems will handle novelty by definition; how they behave in light of novel situations and tasks must be verifiable, at some reasonable level of abstraction, to ensure their safety. Automating explanation generation in AI systems is therefore an important goal [17], and it might be argued that it is necessary for a system to be worthy of being considered general [23].

It is the ability of explanations to be verified that brings them their fundamental value. To be verifiable means that they must be based on knowledge of verifiable causal relationships in the situation, task, or circumstances in question—in other words, they must be *falsifiable*. To be falsifiable they must reference some set of causal relations whose validity is undisputed in the relevant contexts, or easily verifiable.

An important function of explanation that is less often discussed than most others is their use for guiding an autonomous agent's learning; the ability to find explanations for learning failure or success can help uncover how the world works. In this case, to be effective and efficient, explanation generation must be autonomous [17]. Here we examine *goal-directed self-explanation:* the ability of a system to autonomously generate explanations about its own behavior, as well as its acquired knowledge of tasks and environment, under articulated requirements, i.e. explicit goals. A key focus of this work is the use of such explanations as a method for learning (and meta-learning, that is, learning to learn).

The work rests on the argument that explanation generation is a fundamental and necessary process for general self-supervised learning [23]. We look at how explanation generation for this purpose is achieved in the AERA system, and discuss its approach in light of other systems aimed at general intelligence. Thórisson [21] describe a theory of pragmatic understanding that we take as the foundation for our work here. We consider their definition of understanding well-suited for building a theory of explanation generation because it already presents a strong foundation for relating prediction, goal achievement, knowledge acquisition, and explanation to causal reasoning.

The paper is structured as follows: We start with an overview of related work, then we provide some important definitions for the subsequent discussion, which outlines our theory of goal-driven self-explanation generation.

2 Related Work

For reasons of opaqueness, studies on explainable AI have so far primarily focused on artificial neural networks (ANNs), being mostly based on (manually guided) abductive methods that attempt to trace certain outputs to the identification of relevant inputs (cf. [14]). For immediate clarification, this is not what the present

paper is about. In the allocentric methodologies employed in the development of these systems [19], training data, implicit goals, and hand-coded heuristics, are all determined and provided by the developers themselves, a-priori. In this sense, ANN-based systems are no different from standard software applications.

We envision the aims of 'explainable AI' research differently. First and foremost, we recognize that the primary practical application of AI is all sorts of automation, and therefore *autonomous* explanation generation should be a primary goal for explainable AI. In short, the human effort needed to arrive at an explanation should be minimized as far as possible, delegating the explanation generation to the machine. Equally importantly, we see explanation – and its extension into argumentation in general – to be a foundation for any general machine intelligence to grow its knowledge reliably, efficiently and effectively.

We are working exclusively on systems that can generate explanations autonomously, about themselves and their task-environment—i.e. systems that are *self-explaining*. Generally speaking, explanations can vary in their quality. A good explanation eliminates blind spots, clarifies, or highlights that which was obscure before (see section below). Above all, a good explanation observes certain implicit (explicit) constraints and does not break any relevant rules. To do so, it is not enough that an explanation refer to correlational data, it must be based on actual and relevant causal relations. This is because a good explanation must highlight *why* something – whether it be a course of events, situation, or other outcome – *must be the way it is*, rather than some other way [15, 4].

Most sources agree that causal attribution, or identifying underlying causes of a class of (or particular) events or state of affairs, is a vital part of explanation [24, 9, 10, 18, 3].[4] In fact, this is often how explanation is defined. Josephson equates finding possible explanations with finding possible causes [7], and Halper and Pearl claim that explaining a set of events necessitates the acknowledgment of the cause of those events [3]. Miller expands on this, arguing that explanation begins with the cognitive process of identifying causes, followed by a social process of conveying the knowledge acquired by the cognitive process to the intended recipient. As he also points out, causal attribution is a twofold process of inferring the key causes and then selecting a subset of those causes as the most relevant for an explanation [10]. Our approach is somewhat aligned with this view.

Halpern and Pearl [4] define causal explanations using structural equations, for the purpose of determining and conveying an *actual cause* of an explanandum. To accomplish this they assume that all relevant facts are known to said model. What is lacking is a treatment of tasks and goals rather than simply explaining. The assumption of a complete model is also unrealistic, particularly in complex real-world situations. Their work thus leaves much to be desired when it comes to AI, including how such models are autonomously built. This is addressed

[4] Other types of explanation than causal have been proposed. Teleological explanations are explanations focused on utility (to explain by defining the purpose or intent of the thing to be explained [2]). But nowhere nearly all things in need of explaining have intent or utility behind them.

in our AERA system by positioning explanation as the provisioning of missing information structures, making incomplete knowledge a feature, not a bug.

Hilton [6, 5] researched explanations extensively from a psychological perspective. They point out the inherent fallacy in using covariational criteria for causal attribution, as there are numerous examples of events occurring at the same time without one being the cause of the other. Their alternative model of explanation is based on findings in ordinary language where humans make use of contrastives and counterfactuals as criteria for causal attribution. This is also one of the major findings of Miller's survey [10] on explanations: explanations in human conversation most commonly are produced in response to contrastive questions, for instance *"Why did you do A and not B?"* rather than simply *"Why did you do A?"*. Halpern and Pearl [3, 4] also build on this idea, positioning counterfactuals as a way to highlight actual causes.

Palacio et al. went with a broader definition of explanation, arguing that causation is not necessary for all explanation: "An explanation is the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts (by a human consumer)" [14, p. 5]. They further argue that understanding is unique to humans, and therefore explanation from machine to machine is merely verification. We do not agree with either assertion—indeed, we consider it a central task artificial general intelligence research to endow machines with understanding [21], and we see causal relations as central in all explanations (if not explicit, then certainly implicit), because they are the fundamental method for *explanation verification*.

In our view, all explanations of complex tasks with multiple steps and subgoals must be based, in one way or another, on causal relations. We therefore treat causation and causal knowledge as a necessary element in this work.

3 Definitions

Here we give a compact description of key terms used in the following sections, in particular Section 4.

Explainer and Explanation. A process that produces explanations is an 'explainer.' This can be a human, a machine, or some other process which is positioned to serve such a role. An 'explanation' is a compact description outlining some subset of a modelset of the phenomenon that, for whatever reason, is misunderstood, misrepresented, or missing from the phenomenon's modelset. An explanation typically references existing parts of a modelset and presents either a missing piece or highlights errors in it (see Section 4, page 7).

Explanandum and Explainee. 'Explanandum' is that which is to be explained. Given a particular outcome, situation, or turn of events, this can be an anomaly, a missing but necessary relation, or other identified inconsistency that calls for an explanation. 'Explainee' is the particular recipient of an explanation—those to whom the explanation is addressed. This can be the Explaining process itself, a co-located interlocutor, or some future recipient of the information.

Explicit Goal. A 'goal' is a (constant state or steady) state to be achieved. An 'explicit goal' is one which can be described in some representational language that references a knowledge base. An 'active goal' is one which can be thus represented and which may already be pursued—i.e. a goal that has been accepted by an agent of change who is actively pursuing it.

Explainable vs. Interpretable. The terms 'explainable' and 'interpretable' are often used interchangeably in AI, but we see a definitive and important difference between the concepts behind them, based on who exactly is doing the explaining and interpreting. For instance, in work involving artificial neural networks, 'interpretation' is typically an explanation of the mechanisms of the classifier, not of the task or environment for which the system is deployed [8]), and it is the researchers who are doing the interpretation.[5] In contrast, we define *self-explaining AI* as 'AI that is capable of generating valid explanation,' and *interpretable AI* as 'AI that can be interpreted (or explained) by a third party.'

Phenomenon. A phenomenon Φ (process, state of affairs, occurrence) – where W is the world, and $\Phi \subset W$, – is made up of a set of elements, including sub-structures, component processes, whole-part relations, causal relations, or other sub-divisions of Φ $\{\varphi_1 \dots \varphi_n \in \Phi\}$ of various kinds, including relations $\Re_\Phi$ (causal, mereological, positional, episodic, etc.) that couple elements of Φ with each other, and with those of other phenomena.

Complex Task-Environment. We define a 'task-environment' as the tuple of an assigned task and the environment in which the task is to be performed. A 'complex' task-environment is, for all practical purposes, a combination of an assigned task in a particular environment that, for accomplishing the task, requires (a) detection and separation of patterns and sub-patterns with non-trivial causal and part-whole relations, that must be combined with (b) assumptions about high-level logical relations between these (e.g. objects cannot be in to places at once), combined with (c) creation, execution, and monitoring of partial non-linear plans with nested contingency composition, and/or (d) direct application of ampliative[6] reasoning and analogy generation.

Valid Explanation. An explanation $\varepsilon(x, y)$, where x is the explanandum and y is a network of known (causal) relations and patterns relevant to x, can be validated through a process that seeks to uncover inconsistencies in it through the generation of questions that probe y's causal relations relevant to x. To do so the validating process must be able to (a) represent causality, and use this to (b) abduce arguments which "argue for" – or serve as verifiable evidence for – the validity of the explanation. The arguments could also be verified by direct

[5] Providing adequate levels of transparency modern machine learning and AI systems such as reinforcement learners and deep neural networks, with adequate levels of transparency, requires considerable post-hoc effort and skill in interpreting algorithms, and most of the time it is essentially prohibitive due to cost.

[6] Traditionally, 'ampliative reasoning' refers to any process that relies on abduction and induction in any combination to achieve a particular result (cf. [16]); we include (defeasible, non-axiomatic) deduction in that list.

measurement (but is only necessary if the background assumptions on which the evidence rests are not well-verified).

4 Goal-Driven Explanation Generation

We base this work on a theory of pragmatic understanding proposed by Thórisson et al. [21] which uses the concept of a modelset (set of peewee models[7]) for describing a phenomenon, and that can be manipulated through a set of processes for performing four types of tasks, one of which is explanation generation. Given a phenomenon Φ, M_Φ is the modelset intended to capture relevant aspects of the phenomenon; the models ($\{m_1 \dots m_n\} \in M_\Phi$) are information structures intended for (a) *explaining* Φ, (b) *predicting* Φ, (c) producing effective plans for achieving goals with respect to Φ, and (d) (re)creating Φ in any medium (see Section 4, p. 6). For any modelset M_Φ and phenomenon Φ, the closer the information structures as a whole represent key elements (sub-parts) $\varphi_i \in \Phi$ and their couplings $\Re_\Phi$, at any level of detail, the greater the *accuracy* of M with respect to Φ. The more *completely* such a modelset captures all relevant aspects of Φ for achieving any of the four tasks, for any chosen challenge related to Φ, the more *comprehensive* it is. Our theory of goal-driven self-explanation considers explanation generation itself to be *a task* with a particular top-level goal—namely:

> $\mathcal{G}_{top}$ — *The goal of explanation is to improve (or prove) understanding.*

This statement would in itself be a rather shallow if what we mean by 'understanding' was left unexplained; our definition of understanding is exactly this: The more correct – i.e. *comprehensive* and *accurate* – an intelligent agent's modelset M_Φ of Φ is, the better will the agent be said to *understand* phenomenon Φ [21]. An explanation in this view is a concrete action that is intended to verify, evaluate, or increase either the completeness of an agent's models and relations ($Q_{compl}(M_\Phi, \Re_\Phi)$), its accuracy ($Q_{acc}(M_\Phi, \Re_\Phi)$), or both.

As mentioned above (p. 5), the models of a phenomenon's Φ relations $\Re_\Phi$ describe how its elements relate to each other, and to other phenomena. If we partition $\Re_\Phi$ into two disjoint sets, *inward facing* relations $\Re^{in}_\Phi = \Re_\Phi \cap (2^\Phi \times 2^\Phi)$ and *outward facing* relations $\Re^{out}_\Phi = \Re_\Phi \setminus \Re^{in}_\Phi$, an agent whose models are only accurate and complete for $\Re^{in}_\Phi$ understands Φ but not Φ's relation to other phenomena (i.e. its context); an agent whose models are only accurate and complete for $\Re^{out}_\Phi$ understands Φ's relation to other phenomena but will have limited or no understanding of Φ's internals.

A *good explanation* is one that unequivocally demonstrates or verifies understanding of a phenomenon Φ [1], or improves understanding of Φ by affecting the modelset describing the phenomenon in a way that improves the possessor of that modelset's ability to achieve the four tasks related to a phenomenon.

[7] Small models that can be composed into larger modelsets; see e.g. [11, 13].

The explanation generation process involves the skills of identifying (i) the role that the explanation should fulfil, (ii) the relevant patterns and relations that must be referenced for it to serve this role, and (iii) producing a description that meets these requirements (for a particular set of explainees). This is compatible and in line with earlier work on explanation generation (cf. [15, 4]). With the exception of the first skill, to achieve any of these in a complex environment requires information about cause and effect, the knowledge representation capable of supporting the above must, by definition, contain information about the causal structure of Φ.

Generating an explanation calls thus for certain necessary information and must meet certain necessary requirements. More specifically, producing an explanation involves the generation of a *compact description* that references or implicates one or more causal relations that – if not present, or structured differently – would result in a different outcome. The causal relation(s) relevant to the phenomenon that explanation targets limit(s) the possible state space by providing constraints, thus contributing to a particular outcome or situation. The necessary ingredients to produce explanations are, therefore:

- knowledge of causal (and other) relations,
- named entities (and appropriate grammar) for producing this description,
- a fulfillment of a (possibly hypothesized) goal that the explanation is intended to meet.

We hypothesize three classes of purposes – or *subgoals* – that a generated explanation may serve, namely, to highlight or identify the following aspects relevant to an explanandum:

$\mathcal{G}_1$ — Unknown or hidden variables, patterns, or other aspects.
$\mathcal{G}_2$ — Unknown or hidden causal factors and chains.
$\mathcal{G}_3$ — Unknown or hidden errors in background assumptions.

The task of an explainer (explanation-generating process) is to meet the top-level goal that explanation serves, that is, to prove/improve understanding, by meeting one or more of these three subgoals as closely as possible. The explainee can be co-temporal and co-spatial, (as in human realtime dialog), a future receiver of a recorded or written explanation (e.g. instruction manuals), a group of students (as in a classroom), or the explanation-generating process itself (like during learning, when explaining things to oneself for verification of understanding).

Since an explanation serves a purpose, as defined by its subgoal(s), $\mathcal{G}_{1-3}$, we can assume that it may do so on a continuum, from well to badly. The gradient from meeting this goal perfectly, $\mathcal{R}(\varepsilon) = 1$, to not meeting it at all, $\mathcal{R}(\varepsilon) = 0$, describes how well an explanation "hits the spot"—let's call it the explanation's *role fulfillment*, $\mathcal{R}(\varepsilon, \varpi)$, where ϖ is its designated role. And since an explanation could in theory highlight the relevant patterns, causal chains, or background assumptions anywhere from perfectly to not at all, we can define a gradient for this dimension as well, $\varepsilon(P_{rvt}) = [0, 1]$; we call it the *validity* of an

explanation, $\mathcal{V} = \varepsilon(P_{rvt})$. The *value* of a given explanation is then the product of *how well* it meets its goal and how *valid* it is, $v_{pur}(\varepsilon) = \mathcal{R} \times \mathcal{V}$.

We call this an explanation's "*pur* (pure) value" because there is a third factor that could be considered here, that is, how well the explanation fits an explainee agent's A knowledge, $\mathcal{K}(A)$. A 'perfect explanation' is defined as an explanation whose pure validity is at maximum, $v_{pur}(\varepsilon) = 1.0$, and whose compactness could not be greater. The maximum compactness of an explanation ε is in part dictated by this factor, because the more an explainee knows, the more compact can the explanation be made. If an explainer makes incorrect assumptions about the explainee's knowledge – that is, there is misalignment between the explainee's knowledge and the explainer's model of that knowledge – the compactness of the explanation will suffer. We propose to represent this relationship as a match, or overlap, between the constructed explanation's *encoding* and the explainee's ability to unwrap that encoding (in other words, the effort required to decode the information it is intended to carry), that is, $\{\varepsilon_{\Phi} - (\Phi \setminus \mathcal{K}(A))\}$, where Φ is the explanandum, ε_{Φ} is the (encoded) explanation of a particular part of Φ that references both known and unknown information, and $\mathcal{K}(A)$ is the knowledge of the explainee.[8] This, then, may be taken into account when quantifying the value of an explanation.[9]

In a reflective controller, i.e. one that can reflect on its own inner operations, any explanation can become the subject of the agent's own explanation machinery, allowing for the generation of explanations of explanations (like we are doing right here right now). Capacity for this kind of self-explanation can enhance not only an AI system's understanding of its task and environment but also of *itself*. In each case the explanations coming from within the system can be processed by the system for the purpose of further knowledge acquisition [23]. Stated differently, given that the system is a *self-explaining* AI, the better the above explanation generation functions are fulfilled and implemented in the same system, the more trustworthy the system will be, but not only that, it could possibly learn faster and better. Going one step further, a paper by Thórisson argues that autonomous general learning is not possible without some form of explanation-generating mechanisms [23].

5 Explanation Generation in AERA

This section gives a short introduction to how AERA (Autocatalytic Endogenous Reflective Architecture) meets the above requirements for generating explanations [11, 12]. Knowledge in AERA is represented using two main types of information structures, composite states and causal-relational models (CRMs) [22, 13,

[8] For convenience we include, as part of the 'encoding' of an explanation, any references to related but different phenomena intended to better match an explainee's knowledge—that is, to explain something better to a particular explainee, due to their particular knowledge at the time of the explanation generation.

[9] This certainly is a factor in all explanations produced by one human for another. It may not, however, be relevant for self-explanation generation since the meaning of a low-value (or zero-value, i.e. worthless) explanation produced for oneself is undefined.

11]. Composite states capture patterns that an AERA agent can perceive; CRMs capture causal relations by representing causes on the left-hand side and results on the right-hand side. Pattern matching is used to match perceived or desired states to either side. Using these constructs, AERA learns in a self-supervised way by constructing programs on the fly for achieving self-generated goals and sub-goals [20]. The resulting networks of information produce both concrete and hypothetical plans, predictions, and sequences of actions that fulfill set goals.

AERA's capacity for self-explaining comes primarily from two key principles. Firstly, all its knowledge is explicit and compositional in a scale-independent way. This means that both small and large details can be captured with comparable information structures, and that hierarchies of knowledge can also be constructed into modelsets (through combinations of smaller elements). Secondly, because cause-effect relationships are represented directly (also in a relatively scale-free manner), computing the implications of particular actions, and producing appropriate plans for achieving goals, is directly supported.

Finally, the special programming language used to implement these mechanisms in AERA, Replicode [11], makes key parts of the system's operational semantics accessible to itself, allowing it to use explanation to argue *to itself* about which action to take, which options may be better than others, and what particular actions may lead to in comparison to others.

6 Conclusion

Explainability and traceability are key requirements of all mission-critical engineering. With the increasing use of software-controlled systems, complexity rises, and with complexity comes the need for smarter software systems. To be trustworthy, AI must be explainable. With the goal of creating systems with general intelligence, AGI-aspiring systems should not only be explainable, they should be able to explain themselves to their users. But if general intelligence *requires* the ability to explain – if not for any other reason that the sheer amount of possibilities that the physical world presents to anyone who is learning about it from scratch – then such systems, upon having achieved generality in the near or distant future, will already be able to generate good explanations about their own operation and their task-environment. We hope the work in this paper moves us one step closer to this future.

Acknowledgments. This work was supported in part by Cisco Systems, the Icelandic Institute for Intelligent Machines and Reykjavik University.

References

1. Bieger, J., Thórisson, K.R.: Evaluating understanding. In: IJCAI Workshop on Evaluating General-Purpose AI, Melbourne, Australia (2017)
2. Cohen, J.: Teleological explanation. Proceedings of the Aristotelian Society **51**, 255–292 (1950)
3. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach — part I: Causes. Brit. J. Phil. Sci. **56**, 889–911 (2005)

4. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach — Part II: Explanations. Brit. J. Phil. Sci. **56**, 843–847 (2005)
5. Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin **107**(1), 65–81 (1990)
6. Hilton, D.J., Slugoski, B.R.: Knowledge-based causal attribution: The abnormal conditions focus model. Psychological Review **93**(1), 75–88 (1986), `https://doi.org/10.1037/0033-295X.93.1.75`
7. Josephson, J., Josephson, S.: Abductive Inference: Computation, Philosophy, Technology. Computation, Philosophy, Technology, Cambridge University Press (1996)
8. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications **10**(1) (Mar 2019). https://doi.org/10.1038/s41467-019-08987-4
9. Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Sciences **10**(10), 464 – 470 (2006)
10. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences (2017)
11. Nivel, E., Thórisson, K.R.: Replicode: A constructivist programming paradigm and language. Technical Report RUTR-SCS13001, Reykjavik University School of Computer Science (2013)
12. Nivel, E., Thórisson, K.R.: Towards a programming paradigm for control systems with high levels of existential autonomy. In: International Conference on Artificial General Intelligence. pp. 78–87. Springer (2013)
13. Nivel, E., Thórisson, K.R., Steunebrink, B., Dindo, H., Pezzulo, G., Rodríguez, M., Hernández, C., Ognibene, D., Schmidhuber, J., Sanz, R., Helgason, H.P., Chella, A., Jonsson, G.K.: Bounded recursive self-improvement (2013)
14. Palacio, S., Lucieri, A., Munir, M., Hees, J., Ahmed, S., Dengel, A.: Xai handbook: Towards a unified framework for explainable ai (2021)
15. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, New York, NY, USA, 2nd edn. (2009)
16. Psillos, S.: An explorer upon untrodden ground: Peirce on abduction. In: Handbook of the History of Logic, vol. 10, pp. 117–151. Elsevier (2011)
17. Rörbeck, H.: Self-Explaining Artificial Intelligence: On the Requirements for Autonomous Explanation Generation. M.Sc. Thesis, Dept. Comp. Sci., Reykjavik University (2022)
18. Strevens, M.: The causal and unification approaches to explanation unified-causally. Noûs **38**(1), 154–176 (2004)
19. Thórisson, K.R.: A new constructivist AI: From manual construction to self-constructive systems. In Theoretical Foundations of Artificial General Intelligence (Pei Wang and Ben Goertzel, eds.) **4**, 145–171 (2012)
20. Thórisson, K.R.: Seed-programmed autonomous general learning. Proceedings of Machine Learning Research **131**, 32–70 (2020)
21. Thórisson, K.R., Kremelberg, D., Steunebrink, B.R., Nivel, E.: About understanding. In: Proceedings of the International Conference on Artificial General Intelligence. pp. 106–117. Springer-Verlag, New York, NY, USA (2016)
22. Thórisson, K.R., Talbot, A.: Cumulative learning with causal-relational models. In: International Conference on Artificial General Intelligence. pp. 227–237. Springer (2018)
23. Thórisson, K.R.: The 'Explanation Hypothesis' in general self-supervised Learning. Proceedings of Machine Learning Research **159**, 5–27 (2021)
24. Woodward, J.: Making things happen: A theory of causal explanation. Oxford university press (2005)